Integrating peer assessment in teaching: A subject specialist perspective

Vasi van Deventer, Department of Psychology, University of South Africa

Abstract

This paper describes how a research methodology course became structured around a peer assessment procedure. The peer assessment procedure was designed to form an integral part of the course in operationalising certain aspects of the learning model on which the course was founded. This kind of design process requires a substantial philosophical and theoretical import. One has to consider various contexts of teaching, survey different instructional philosophies and come up with a well defined learning model to establish a meaningful integration of peer assessment and course work. In this paper I provide a brief outline of the philosophical issues which guided the definition of our learning model. I explain the learning model in some detail to trace the roots of peer assessment before describing the logistics of the assessment procedure. The paper concludes with a discussion of empirical information obtained from the peer assessment procedure.

At the beginning

Peer assessment is a technique teachers can use in a variety of ways without having to ground it in a particular teaching approach. And this is exactly how I first stumbled across peer assessment. It happened by accident. About ten years ago I had to teach 80 post graduate students how the neuro-physiology of memory relates to certain underlying philosophical themes. Due to the complexity of the work I offered the material as a series of excercises culminating in a journal type essay on the topic. Purely as an after thought I had them peer assess these essays to learn from each other, and also because the idea of having to grade 80 full length essays did not really appeal to me. But the process worked so well that I introduced it in a second course where students had to write an essay on the psychology of religion and the body-mind problem. Much to my amazement I realised that I could use the same assessment criteria for both essays. This taught me an important lesson, namely that one can devise assessment criteria that are generic to a learning experience and not dependent on the specific content of the experience. I did not have the terminology then, and did not really know what I was doing, but I intuitively understood that students who master the generic dimensions of a learning experience were much better off than those who study content only. It was this idea which guided us when three years ago we decided to revise our post graduate course in research methodology. I was determined that the course should be organised around the generic dimensions of the research process, that we could use peer assessment as a formative measure during the course, and that we should design the course to allow for yearly updates to accommodate the experience we gain during the process. Two factors, namely empirical information gathered during teaching the course and increased debate around instructional design issues in our university have contributed to peer assessment becoming a properly founded and integrated technique in our teaching approach.

This paper provides a brief outline of factors that guided the revision of our research methodology course. These factors necessitated a paradigm shift in our teaching approach and required us to define a learning model which could guide our activities. I sketch three instructional design paradigms to show where our learning model comes from and where it fits in philosophically. I then describe our learning model in some detail to trace the roots of peer assessment before explaining the logistics of the peer assessment procedure we implemented. The paper concludes with a discussion of empirical information obtained from the peer assessment procedure.

The need for a paradigm shift

Various factors compelled us to change direct in our research methodology teaching, factors that came from different contexts, namely a global context, a national context and the context of the subject field itself.

The global context: Learning individuals in a world of praxis

Two of the most pronounced phenomena of the era we live in are the rising power of the individual and the fact of constant change. We seem to have resigned ourselves to the idea that constant change has replaced the stability of ultimate truth as the fundamental condition of our world. We see ordinary individuals constituting the millions of nodes of worldwide webs, bypassing the borders and limitations of their national geographies. According to Davidson and Ross-Mogg (1998) the power of the individual is fast exceeding the power of overarching governmental bodies to such an extent that the days of the nation state are numbered, and they are numbered in decades not centuries. The individualised world of constant change molds a new kind of citizen, citizens we have come to define as learning individuals (Schwahn & Spady, 1998). Learning individuals are not subject to unshakeable truths, they are not subject to being taught in ways which rob them of their individual agencies in learning. Learning individuals demand a change in paradigm. They want new models of learning. In South Africa the Green Paper on skills development strategy for changing global conditions (Department of labour, 1997) says learning individuals need applied competence. It describes applied competence in terms of three kinds of competence, namely practical, foundational and reflexive competence. Practical competence is seen as the demonstrated ability to perform a set of tasks. Foundational competence reflects the demonstrated understanding of what one and others are doing, and why, and reflexive competence is the demonstrated ability to integrate or connect one's performance with one's understanding of those performances such that one learns from one's actions and are able to adapt to changes and unforeseen circumstances.

All of this boils down to the fact that we live in a world of praxis, a world in which actions and reflection about those actions are of paramount importance (Grundy, 1987). It is a real world, not a hypothetical / theoretical world. It is a world of interaction, a social and cultural world. It is not a world of absolutes, but a world of meaning making.

The national contex: Deficiencies in researh methodology teaching

The wakes of global change are felt strongly in South Africa. We are a society experiencing accelerated transition and there is no inertia to buffer its citizens against the effects of global change. This is the kind of environment in which social scientists have a critical role to play. The Centre for Science Development (CSD, 1997), a research funding arm of the Human Sciences Research Council (HSRC), conducted a national audit of social science research training at South African universities and technicons. They found the traditional ways of research training questionable in light of the changing context of higher education. For example, South African scientists are confronted with an increasingly wide range of issues requiring scientific investigation, implicating a wide range of methodologies for which our young scientists are simply not prepared. Tertiary institutions are slow to take cognisance of changing educational content and practice. Also, they have to deal with increasing limitations on resources while handling large numbers of students. The CSD findings suggest more inter-disciplinary and inter-institutional co-operation in research methodology teaching, and more real life material based and experiential approaches to learning.

Our own experience supports the concern that traditional research training does not equip students for practical realities. Students pass traditional research courses but still lack the skills to plan their Masters degree research projects. In examinations we often find that students who are quite capable of answering complex theoretical methodological questions cannot relate their knowledge to practical reality.

The subject context: New ideas about knowledge

The third context which prompted drastic revision of our research course concerned developments in the subject field itself. Hermeneutic, critical and action based methodologies supplement the logical positivistic approaches traditionally taught in social science research courses. Academic publications reflect increased references to these new methodologies, and even in our own department (Psychology) we experience the turn to new ways of thinking about knowledge. By the mid 1990's we had already seen the completion of several doctoral studies, either questioning conventional methodologies or themselves using nontraditional research approaches. These developments are not isolated from a more general global zeitgeist. The newer methodologies seriously question the notion of absolute truth and also call for a much greater reliance on the personal agency of the social scientist.

The effect: A different kind of scientist trained in a different

way

We did not really have a clear understanding of the contextual issues described above when we began contemplating changes to our research methodology course (for example, the Green Paper on skills development and the CSD audit of research methodology training had not been available at the time), but we were well aware of a general undercurrent nudging us in a different direction. We held a workshop involving subject specialists on the teaching team and an instructional designer from our Bureau for University Teaching (BUT), to form an idea of the kind of social scientist to be delivered by the end of the course. The workshop produced the following descriptions of the kind of social scientist needed in South Africa:

- Our social scientists should be self-actualising and reflexive. They should foster a criticality which enables them to participate in self- and societal transformation by critiquing and reconstructing received knowledge.
- Our social scientists should be able to identify and participate in socially significant projects. They should be able to conduct research in collaboration with communities, aiming to furthering social justice.
- Our social scientists should uphold emancipatory aims, emphasising open and democratic relationships, fostering power-sharing and participatory control (especially in research projects).
- Our social scientists should understand scientific facts as value-laden and context specific.
- Our social scientists should act as agents in the deconstruction of the discipline of research methodology, that is, they should understand that Africanisation requires a deconstruction that goes beyond adding African relevant examples to traditional subject material.

We realised that these social science researchers were our new learners and that our traditional teaching approach could not effect the necessary changes. Our teaching was not organised around principles of self-actualisation and reflexion. There was nothing emancipatory and democratic about the way in which we engaged our students. Power-sharing and participatory control were things we preached but never practised.

The meaning of a paradigm shift

The fields of research methodology and instructional design share a great deal. Both areas are concerned with the nature of knowledge and ways of acquiring knowledge, and both areas are marked by diverse positions concerning theory and application (Richey, 1995). Both areas seem to unpack their philosophies in terms of traditionalist, hermeneutic and critical ideals.

The traditionalist, hermeneutic and critical perspectives

According to the traditionalist perspective information and skills exist in the world "out there". Facts can be determined positively and objectively, and can be asserted as public truths. Facts can be distinguished from believes in the sense of being scientifically determined and/or rationally deduced from prior facts. To gain access to ultimate truth the researcher has to be objective about the researched. A fundamental split exists between the researcher and the researched, and also between knowing and doing. Factual knowledge is divorced from actions. Knowledge serves to guide actions, but actions do not lead to knowledge.

The hermeneutic perspective sees knowledge as socially constructed, and historically and culturally specific. Subjective understandings are important as the researcher cannot divorce herself from the researched context. Actions and factual information inform each other. Actions are guided by factual knowledge, and knowledge comes through actions.

The critical perspective goes one step further: Not only is knowledge socially constructed, it also has political interests. The critical perspective encourages an acute awareness of the relationship between knowledge and power. It is serious about the dialectic between knowledge and the agency of the knower. Knowledge is validated in praxis (in and through doing things) within specific social and political contexts.

Instructional and learning implications of the different perspectives

The distinction between the traditionalist, hermeneutic and critical perspectives serves as a useful background for contrasting different aspects of instruction and learning processes, aspects such as curriculum, goals of education, learning theory, the teacher-learner relationship, and assessment.

- Curriculum: The traditional perspective sees the curriculum as a product, consisting of a particular content which governs teaching inputs and learning outputs. The hermeneutic perspective views the curriculum as practice in the sense that teaching input is based on the teacher's professional judgement and the learner's understanding. From a critical point of view the curriculum appears as a field of praxis in which teacher and learner work together as social agents in transforming institutions and society.
- Goals of education: The traditionalist teacher aims to equip students with the necessary knowledge

and skills. Hermeneutic teachers try to produce self-actualising, reflective "educated people", and those following a critical approach aim to produce self-actualising, reflective "educated people" with a criticality which enables participation in self and societal transformation.

- Learning theory: From the traditionalist perspective the learner has a deficit, and the deficit needs to be addressed. Learning follows behavioural principles. The hermeneutic approach is constructionist and interactionist. The learner builds cognitive structure through interaction. The critical theorist sees learning as a social constructionist and interactionist process in which the learner reconstructs his knowledge self-reflectively in an attempt to move beyond subjective understandings.
- The teacher-learner relationship: In a traditionalist environment the teacher is an authority figure who controls the learning process in a hierarchical relationship. In this relationship status equals power. The hermeneutic teacher is a lender who progressively yields control of the learning process to learners, within a mentoring relationship. In this case status and power are based on merit. From a critical perspective the teacher is a co-ordinator with emancipatory aims. She emphasises commonality of concerns within an open and democratic relationship. The teacher-learner relationship is one of power-sharing and participatory control.
- Assessment: The traditionalist presupposes public truths which can be assessed objectively. Content, and those cognitive skills required to handle the content, are assessed to determine whether the deficit with which the learner arrived at the course has been wiped out. Assessment is authoritarian and relies on external standards. The hermeneutic assessor understands that knowledge is embedded in historical, situational and cultural contexts. Knowledge has relevance and harbours implications. Personal experience contributes to knowledge, making knowledge more like wisdom than pure objective fact. Standards to judge knowledge by can at best be negotiated. The critical theorist believes in the politics of knowledge, and puts a high premium on the ability to act. The level of agency required for particular actions, the effectiveness of such actions, and the quality of reflexive judgement are important indicators of knowledge. The judgement of this kind of knowledge requires internalised standards, imposed by self and peer assessment.

The nature and implications of the traditionalist, hermeneutic and critical perspectives are perhaps best summarised in terms of our geometric learning field model.

The geometric learning field model

We borrowed the basic principles of our geometric learning field model (GLFM) from Einstein's general theory of relativity. This theory is about gravity. Einstein represented gravity in geometric terms. The theory states that a large mass makes a dent in space-time. The most important point is that the amount of mass is equivalent to the size of the space-time dent. In the geometric learning field model knowledge mass replaces physical mass and space-time equals a learning field. In other words, the geometric learning field model is about knowledge masses embedded in a learning field. Any source of information forms a knowledge mass. Thus teachers, learners, books, journal articles, etcetera, all form knowledge masses in a learning field. Every knowledge mass leaves its dent in the learning field, rendering a field that is not flat, but a landscape filled with valleys and hills.

In a traditional teaching domain the learning field is dominated by the teacher's knowledge mass. The teacher represents a body of absolute truths, which is the knowledge mass of the discipline in question. Small student masses are pulled into orbit around the larger teacher mass. Here teaching aims to slow a student’s orbital velocity, finally to make him fall into the teacher mass to become subjected to the discipline's knowledge mass. The hermeneutic teacher, on the other hand, aims at speeding up students' orbital velocities to help them escape the pull of the teacher's knowledge mass. Yet, in both cases the teaching discourse concentrates on knowledge masses, and ignores the learning field. It requires a critical perspective to shift the teaching discourse from knowledge mass to the curvature of the learning field

A learning field is a field of information, and the way it curves into valleys and hills represents knowledge in the field. The classical distinction between data, information and knowledge provides a useful way to describe the learning field. One can begin by picturing the learning field as a flat sheet consisting of billions of bits of data, and then imagine an indent appearing in the flatness of the sheet. The appearance of the indent signifies a source of information. All data in range of the information source, ie all data affected by the indent, are aligned with this source and become information. The size of the indent caused by the information source is equivalent to the mass of knowledge associated with the source. The degree of curvature occurring in the learning field is equivalent to the mass of knowledge in the field. But here one strikes a difficulty. Because one always finds oneself within the learning field one cannot see how the field curves up and down into valleys and hills. In other words, we cannot perceive knowledge - what we perceive is information. Yet knowledge is embedded in the information we perceive, represented in the varying densities of information. The information field in which one finds oneself is not uniform and featureless. The varying densities, which run like wrinkles through the information field, represent the curvature of the field. Knowledge is hidden in these varying densities, which represent the curvature. The curvature of the information field (ie the knowledge in information) defines the reality of information. We cannot see the curvature. We cannot see reality. Yet, the curvature separates what is practically possible from what is practically impossible. It leaves its traces like wrinkles in information, wrinkles we grasp as knowledge.

The learning field model leads to interesting conclusions. From the point of view of any information source (ie learners, teachers, books, etc) the learning field consists of information only. Both data and knowledge are abstractions beyond information. The information source cannot see data for what it really is because the source is always already embodied as a dent in the learning field. Making a dent in the data field aligns data with the source of information. All date in range of the source becomes information for the source. Knowledge also escapes the source. The information source also has no direct perception of knowledge. Knowledge is the information source’s reality, the source’s understanding of what the world is really like. Knowledge constitutes a reality principle, differentiating what is really possible from what may be possible; always hidden in information, never showing itself in its naked truth. Although knowledge belongs to sources of information (all sources of information has a knowledge mass) knowledge is always also interpersonal and multicultural. The curvature at a particular point in the learning field is always a function of all the different dents (knowledge masses) in the point’s vicinity. Knowledge as curvature expresses the interpersonal nature of knowledge. But this does not mean that knowledge is pure objective fact. A number of factors impact on how individuals see the world, factors such as personal beliefs, values and attitudes. There is a parallel in Einstein's theory which allows us to incorporate these kinds of additional factors as part of the knowledge mass. Einstein's calculations show space-time curvature to be equivalent to not only physical mass but to physical mass plus pressures internal to the mass. Internal pressures originate from the tendency of a mass to compact in on itself under its own weight. The equivalent position in our learning field model would be that a dent in the learning field is equivalent to a mass of factual knowledge plus those internal pressures caused by factors such as personal believes, values and attitudes. When learners and teachers find themselves in the learning field each contributes to the curvature of the field (ie the knowledge of the field) with a knowledge mass which harbours its own internal pressures in the form of personal beliefs, values and attitudes.

The process of learning is a process of meaning making, a process of making things fit. It is a process of creating knowledge out of information. Learning means to modify the curve of the information field. To illustrate what this means: imagine having to construct a triangle with each of its angles equal to 90 degrees. This is impossible on a flat surface but easy to do on the curved surface of a sphere. If we cut out the segment of the surface of the sphere outlined by the triangle in question and flatten the segment on a flat surface we land up with a piece of crumbled paper. The creases and folds in the paper correspond with varying densities in an information field. These varying densities contain the amount of curvature necessary to construct the triangle. The meaning making, the attempt to make things fit, requires one to manipulate information such that it acquires a density distribution that corresponds with a particular curvature. However, although the curvature prescribes the density distribution it does not represent an absolute and ultimate truth. The curvature of the learning field is not an everlasting constant. It changes when a knowledge mass is modified (for example when a learner acquires new insights, or experiences a change of attitude, etc), or when additional knowledge masses appear on the scene, or when existing knowledge masses disappear from the scene. In other words, knowledge does not reflect an ultimate and fixed reality. The geometric learning field model assumes that reality and truth are constructed among various different knowledge masses. Learning is not merely a passive process of getting to know reality. It is also to be actively involved in reconstructing reality. This kind of learning happens in learning experiences.

Learning experiences and the geometric learning field model

Learning experiences are marked by three moments in the geometric learning field model. Learners have to engage the learning field curvature to establish the limits of practicality, they have to take cognisance of other knowledge masses to gain a theoretical understanding of the curvature, and they have to track how their own knowledge masses are affected by their practical and theoretical engagements. Thus learners have to develop three kinds of competence, namely practical, foundational and reflexive competence.

Practical competence is required because a learning experience happens in a learning field, and is governed by the curvature of this field. The curvature acts as a reality principle separating the practical from the impractical, thus guiding the development of appropriate performance. The curvature does not prescribe particular actions. Appropriate performance develops through the learner being confronted with the practicality / impracticality of particular actions. Foundational competence is required for judging the curvature of the learning field. Learners with little foundational competence judge the curvature from their own perspectives only, and find it difficult to foresee the practicality / impracticality of their actions. Proper judgement of the curvature requires cognisance of other knowledge masses. These masses could be teachers, peers, or books - any source of information is a knowledge mass with potential impact on the curvature at the point of learning. Thus foundational competence serves to increase the quality of performance, but the ability to distinguish different levels of quality (ie the standard of quality obtained) requires reflexive competence. Reflexive competence involves the reconstitution of the learner's knowledge mass relative to other knowledge masses. To form a clear idea of this concept one has to keep in mind that there is an equivalence between curvature and mass. One also has to remember that the curvature at the point of learning is a function of all the knowledge masses in range of this point. In the learning experience the learner engages a curvature which results from a number of knowledge masses, including his/her own knowledge mass. The curvature of the learner's knowledge mass plays a definite role but is modified by the presence of the other knowledge masses. Mastering the curvature of the learning point means the learner's knowledge mass is modified to approximate the curvature. Reflexive competence is the ability to take cognisance of such modification. It is the ability to contrast different actions and to indicate why certain actions are more appropriate than others. It requires the internalisation of an external standard, thus constituting a personal agency within the learning experience.

This concludes the description of the geometric learning field model. Let’s turn to peer assessment now, to trace its roots in this model.

A foundation for peer assessment

As remarked at the beginning of this paper, peer assessment is neither a philosophy nor constituted by the presuppositions of a particular philosophy. It is merely a technique, a method of assessing someone. However, it can be used to express or to operationalise certain philosophical ideas, in which case it gains a definite philosophical import, as is the case with the geometric learning field model. This model accommodates certain shifts in philosophy. For example, it removes the learner from the orbit of the teacher and puts the teacher in the learning field as one knowledge mass amongst others.

Although the teacher's knowledge mass looks like any other knowledge mass from the learner's point of view the teacher's knowledge mass is, in fact, a very important mass because it determines the generic dimensions of the learning field curvature at the point of learning. These are the dimensions along which learners verify the practicality of their actions. These ultimately form the criteria for judging the quality of actions

In other words, it swops a teaching model for a learning model. It also situates the point of learning in a learning field which is necessarily interpersonal, thus undermining the idea that learning is a solely individualised affair. Yet, it puts the responsibility for learning in the hands of the learner. There is no authoritative teacher telling right from wrong. Learners have to find answers for themselves by verifying the practicality of their actions and by tracking the resulting transformations of their knowledge masses. They have to position themselves relative to other knowledge masses in the field to determine the standards of their actions and establish personal agency.

Peer assessment offers itself as a useful technique to operationalise these ideas. Our peer assessment technique involves three phases, namely the learner producing a product in a learning experience, the learner acting as assessor of his own and his peers' products, and thirdly, the learner re-assessing and modifying his own product. According to the geometric learning field model it is vital that learners have to produce the product themselves before they act as assessors of the product. In having to assess their own products as well as those of their peers self-assessment forms part of the peer assessment process. This allows their peers' products to illustrate potential positions which learners themselves could have taken up in producing the product. In other words, peer assessment always involves self-reference, allowing learners to establish standards for their own work. When learners act as assessors they use dimensions that are generic to the learning experience to judge various aspects of the product. Although these are the generic dimensions of the curvature at the point of learning they do not set standards but merely constitute a platform for comparison. Standards are expressed relative to average group performance. And these standards become internalised when learners have to re-assess and modify their own work in the face of feedback from their peers. Having to justify their actions help them to constitute personal agency relative to the learning experience.

The logistics of the peer assessment procedure

In our case the products submitted for peer assessment are written essays. Although the essay formats differ for the four courses in which we use peer assessment, assessment procedures are similar in all instances. Of the courses currently using peer assessment the course in research methodology provides the best footprint for peer assessment because it was designed around this procedure. This course produces two kinds of essay, namely a research proposal and a research report in the form of a journal article. The assessment procedure involves submitting a series of three assignments, namely the original essay (first assignment), essay evaluations (second assignment), and essay re-evaluation (third assignment). (See Figure 1).

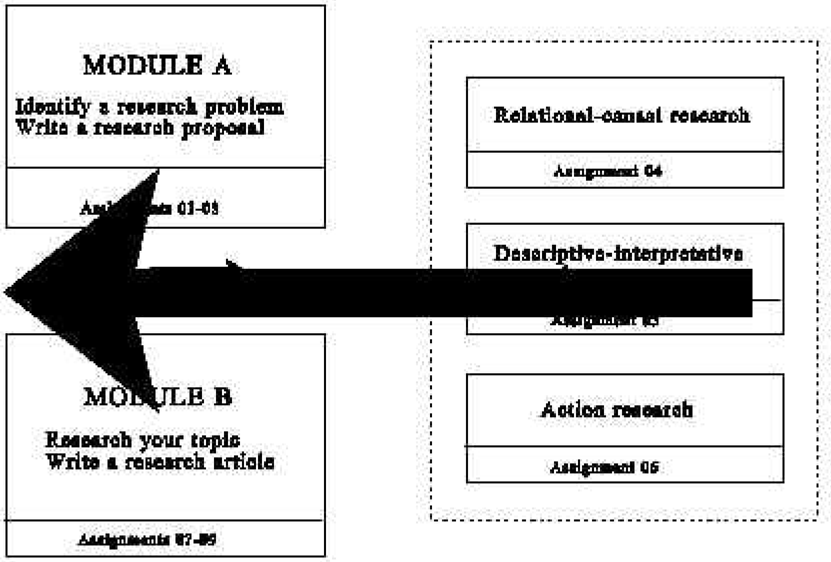

Figure 1: The research methodology course outline

In the research methodology course Assignments 01, 02 and 03 incorporate the peer assessment of a research proposal, and Assignments 07, 08 and 09 deal with the peer assessment of a research report in the form of a journal article. Assignments 04, 05 and 06 are assignments concerning foundational knowledge about research techniques and are not submitted for peer assessment. Each student is required to submit three copies of his/her essay (the first assignment) and keep the original for self assessment. A peer evaluation computer programme (known as Auntie PEP) is used to assign essays to evaluators. During the evaluation phase (ie for the second assignment in the series) each student is required to assess four essays, those of three peers and his/her own essay. (Figure 2 lists the logistical steps in some detail).

Figure 2: The logistics of the peer assessment procedure

- Each student submits three copies of his/her essay as Assignment 01
- Submissions are registered with a computer programme called PEP (Peer Evaluation Programme)
- PEP randomly assigns essays for evaluation - three essays per evaluator
- Each evaluator receives an evaluation schedule, evaluation rating sheets and three different essays
- Each evaluator evaluates four essays (own essay plus three others) and return these as Assignment 02
- Evaluation submissions and marks are registered with PEP
- PEP computes group statistics, assign essay credits (credits received for Assignment 01) and evaluation credits (credits received for Assignment 02) and compiles individualized reports
- Each student receives three evaluated copies of his/her essay as well as an individualized PEP report
- Students re-evaluate their essays in the light of the feedback they received and submit their re-evaluations as Assignment 03.

Students receive an assessment schedule and assessment sheets to complete their assessments. The assessment schedule explains the aspects to be evaluated and provides rating scales and rating criteria for assessing these different aspects (See Figure 3 for an example of a rating description). An essay's ratings are captured on a rating sheet (See Figure 4).

Figure 3: An example of a rating description

RELEVANCE OF EXPECTED FINDINGS: One should also reflect on the relevance of one's research. Who could benefit from one's findings, and in what way could they be expected to benefit? For example, one's findings could benefit scientific theory (progress in knowledge) , or specific individuals, or groups of individuals, or society in general.

Relevance of expected findings

Rate 0: if the relevance of the expected findings is not mentioned

Rate 1: if the relevance of the expected findings is mentioned but not discussed
Rate 2: if the relevance of the expected findings is indicated and discussed but if the discussion is not entirely appropriate in terms of the research variables
Rate 3: if the relevance of the expected findings is indicated and discussed and if the discussion is appropriate in terms of the research variables.

Figure 4: Example of an assessment sheet

EVALUATION SHEET

PROPOSAL NUMBER (Author's student number):

EVALUATOR NUMBER (Own student number):

ALLOCATION OF TOTAL MARK: [/62] = [%] rounded to [%]

The meaningfulness of the title	0 1 2 [ ]
The appropriateness of the title	0 1 2 [ ]
The reason(s) for submitting the proposal	0 1 2 [ ]
Addressing the research problem	0 1 2 [ ]
The research problem and relevant literature	0 1 2 3 [ ]
Formulating hypotheses or statements of intent	0 1 2 3 [ ]
Variables	0 1 2 3 [ ]
Observation categories or levels of variables	0 1 2 3 [ ]
Relevancy of expected findings	0 1 2 3 [ ]
Organisation of information	0 1 2 [ ]
Structure of the presentation	0 1 2 3 [ ]
Style of writing	0 1 2 [ ]
Tone of writing	0 1 2 [ ]

GENERAL REMARKS

Write down any comments and suggestions you may have to offer. The idea is not to be authoritative but rather to open up communication channels. In other words, if the researcher were to be present what would you have said to him/her? Do you agree with the basic design? Then say so! If you do not agree how do you think the design could be changed? Do you find the topic interesting? Does it stimulate new ideas? If so, what are these ideas?

The PEP programme is used to analyse the peer assessments statistically, and to calculate essay and evaluation credits. The programme also compiles an individualised report for each student (See Figure 5). Due to the possibility that students could perceive the statistical procedure as a form of external authority, which they cannot control, a fun personality was created for the computer programme to help students accept the programme on human terms (See Figure 6).

Figure 5: An example of an individualized PEP report

```
PEP REPORT FOR AUTHOR: 3066-446-1
CREDITS FOR ESSAY: 50 Essay grade: B (Good)
CREDITS FOR EVALUATION: 30 Evaluation grade: A (Very Good)

----------------------------------------------------------------------------------------------------------------------------

MARK SUMMARY

Mark self: 75 Fair evaluation: Mark accepted
Mark one: 97 Over evaluation: Mark rejected
Mark two: -1 No mark received as yet
Mark three: 77 Fair evaluation: Mark accepted

FINAL MARK: 76 Group mean: 68 Group standard deviation: 13.59

ESSAY REPORT
```

> Two factors were taken into account in the calculation of essay credits and grades, namely the position of the final essay mark relative to the group mean and the extent to which evaluators could agree about the essay mark.
>
> Congratulations! According to your peers you submitted an essay of fairly high standard. They awarded your essay a B grade for overall quality.
> The mark you earned was above average. It put you in the upper 40% of your class. For this level of performance you received 40 credits.
> There was a fair level of agreement about your mark. However, although the evaluators found your essay fairly well structured and your style of writing mostly clear there might be room for improvement. Read the text again to see where you could improve on the structure of your essay and on your style of writing. Despite the criticism this level of evaluator agreement still earned you an additional 10 credits.
>
> EVALUATION REPORT
>
> The quality of evaluations was determined on the grounds of consistency and congruency. Consistency involved the reliability of ratings and congruency concerned the standards applied in ratings.
> Congratulations, you managed exceptionally well in evaluating your peers. Auntie PEP awarded your evaluations an A grade for overall quality.
> You maintained a very high level of consistency across your evaluations. This level of reliability earned you 25 credits.
> You achieved a high standard in your own essay but you applied an average standard when evaluating your peers' essays. It seemed you expected slightly less of your peers than what you required of yourself. This meant you were not entirely congruent in how you applied standards. This level of congruency earned you only 5 additional credits.

Figure 6: A personality for the peer evaluation programme

We introduced the peer evaluation programme as Auntie PEP to allow students to relate to the programme as a separate entity on the lecturing team. The idea behind this thinking was to divert authority from the lecturer's position to a fictional character. But, although we wanted students to relate to Auntie PEP as an equal they were also to respect her. Thus she was introduced as follows: Meet Auntie PEP. Auntie PEP is our Peer Evaluation Programme. She lives in Stienie Scheeper's computer and she watches over every step you take - ruthless when it comes to slashing credits! So be warned: this one is a real mean bird!!

The peer assessment procedure was designed to piggyback on the existing assignment submission system at the University of South Africa (a distance teaching institution). Administrative departments handling assignment submissions and marks are totally unaware of the peer assessment procedure. The peer assessment side of the procedure is handled by an administrative support person who is part of the course team. In other words, as far as peer assessment is concerned the course team encapsulates the procedure. The peer assessment procedure is entirely transparent for any person who handles these assignments but is not a member of the course team. Piggybacking on existing procedures also makes it easier to interface the peer assessment system with electronic assignment submission procedures currently coming on line at the university. The idea is to completely automate peer assessment for students who prefer to submit their work electronically.

Analysis in peer evaluation

When we designed the research course around peer assessment we were focussing on realising certain aspects of our teaching approach, hoping that the reflexive activities of peer assessment would have good formative value. The procedure's potential for providing trustworthy marks was of less concern. However, we realised that an analysis of the quantitative information to our disposal could help us improve the system. This led to an ad hoc investigation, the findings of which are discussed below. However, instead of the usual statistical format I choose a more informal readable approach, involving minimal reference to statistical concepts. The reader only needs to keep in mind that the correlation coefficients in question are expressed on a scale varying form -1 to +1, that positive correlations are the only ones really making sense in this context, and that coefficients above 0,20 are considered significant and those above 0,25 are seen as highly significant. Coefficients above 0,30 are viewed as really powerful / meaningful correlations (Cohen, 1992).

Before we get to the numbers, though, there is another matter that needs to be mentioned briefly because it sketches the background against which analysis of peer evaluation procedures takes place. This matter concerns circularity, closed groups and standards.

Circular processes in closed groups and the question of

standards

Any peer evaluation procedure which uses the individuals who produce material for assessment also to act as assessors thereof poses a fundamental complication namely that the system closes in on itself and that its processes become circular. No matter how one breaches the system one can never be sure to escape the circularity of processes operating within it. The value assigned to an essay (i.e. the ratings it receives) is a function of both the quality of the essay and the quality of assessment. In other words, if an essay receives a good rating one cannot be sure whether this was due to the quality of the essay, or whether a poor assessor over-estimated the quality of a mediocre essay.

Whether one sees the complication of circular processes as a problem depends on one's philosophical approach and one's perception of the nature of truth. In a critical constructionist model closed group circularity should not pose a problem because one would accept that truth is constructed through the interactions between production and peer assessment. This, after all, is how the tightly closed groupings of academia maintain their standards. But when it comes to learning there is a popular belief in teacher omnipotence, embracing a world where standards are based on absolute truth, the truth of the teacher that often manifests itself as an externally introduced criterion. The general belief is that groups in training are unable to formulate proper standards and that therefore standards should be imposed from outside. In the geometric learning field model this would mean small student masses captured in orbits around a massive teacher knowledge mass.. However, if the belief in externally introduced standards is valid reflexive competence would be difficult to realise. It would be difficult to realise learner independency and to build learner agency in a context where the learner's ability to adapt to changes and to handle unforeseen / unexpected circumstances are measured against externally introduced authoritative criteria.

In what follows information is presented to support the view that groups can constitute standards which would be judged proper by external authorities. This view is based on the fact that marks afforded by peer evaluation (a group standard) correlate significantly with teacher afforded examination marks (an external standard).

The possibility of self-imposed standards

When we first incorporated peer assessment in our courses we expected the circularity of closed group processes to average out in a random distribution of marks. We believed that a group would not be able to assess essay quality in a meaningful way because essays would be subjected to any combination of good and poor evaluators. First results supported this expectation: there was very little difference between within essay variance (the assessments of a particular essay) and between essay variance (the assessments of different essays). For example. the average fluctuation associated with any particular essay's assessments was 10,93, whereas the fluctuation associated with the marks assigned to different essays was 11,94 for a group of 358 students. We confirmed this trend for a second group of 366 students. Here the corresponding values were 12,77 and 13,18. Thus, on average the assessments of any particular essay varied as much as did the assessments for different essays.

However, the similarity of within essay variance and between essay variance did not mean that all essays averaged out on the same mark. In other words an essay could be afforded the average of its assessments and these average marks would form a distribution. The essay marks formed a rank order. It was clear that the ranking did not occur on a random basis because the assessments of the essays were highly correlated. In other words, if an essay received a high mark by one evaluator it was likely to get a high marks by its other assessors. Table 1 shows these correlations for the 1997 (to the left of the main diagonal) and the 1998 groups (to the right of the main diagonal). (The number of pairs differed slightly for each correlation because not all assessments were submitted).

Table 1: Intercorrelations of peer and self assessments

	Assessment 1	Assessment 2	Assessment 3	Self Assessment
Assessment 1		0,50 (n = 319)	0,44 (n = 320)	0,38 (n = 308)
Assessment 2	0,43 (n = 300)		0,49 (n = 322)	0,41 (n = 313)
Assessment 3	0,49 (n = 293)	0,45 (n = 286)		0,45 (n = 312)
Self Assessment	0,36 (n = 292)	0,36 (n = 283)	0,35 (n = 276)	

The significance of these correlations showed that the peer assessment procedure did not provide random results. But there was a problem with the reliability of essay marks. Statistically speaking an essay's mark could pop up virtually anywhere in the distribution of essay average marks. The interval within which one could expect an essay's mark to fall was quite substantial due to the large average deviation associated with inter-essay assessments. One could expect large overlaps of the intervals associated with different essays. Thus, although essays were put in a particular rank order by their peers one could not be sure about the validity of the ranks assigned to individual essays. This problem led us to investigate two strategies aimed at reducing the amount of deviation associated with essay assessments, namely the best bet strategy (BBS) and the evaluation deviation strategy (EDS). Whereas the best bet strategy relies on a minimum of information, the evaluation deviation strategy tries to maximise usable information.

The best bet strategy holds the point of view that marks which are numerically more similar gives the best bet of an essay's real value (We prefer the term *real value* to *true value* in an effort to avoid the idea that an essay has properties which are unshakably true. We see *real value* is an assigned property reflecting the value which assessors are most likely to agree on.). Thus, for each essay the author's self assessment is ignored and from the remaining three peer assessments the two marks differing least in numerical value are chosen. These two marks becomes the essay's best bet marks. The strategy guarantees a lower ratio of within essay variance to between essay variance. For example, the strategy succeeded in lowering the average standard deviation of peer assessments (excluding self assessments) from 13,149 to 4,838 in a group of 364 students.

Although the best bet strategy is a simple and useful technique to rank candidates it has statistical and psychological drawbacks. Forcing essays to slot into a particular order is no guarantee that individual essays are assigned their proper ranks.. This kind of reliability is no guarantee for validity. Furthermore, the procedure excludes fifty percent of available data. It also disempowers students by excluding self assessment marks, and there is, in fact, no reason why self assessment marks should not be taken into account, as the correlation between self assessment and the average of peer assessments turns out to be 0,37. The second approach we investigated, namely the evaluation deviation strategy, tries to maximise usable information, and includes self assessment data. According to the evaluation deviation strategy any assessment should be seen as contributing to constructing the value attached to an essay. But this does not mean that all contributions (ie assessments) are equally useful. The more useful contributions are the more they should be inclined to cluster, due to being governed by a shared ideal. On the other hand one should keep in mind that not all essays are equally easy to assess. Thus not all essays allow assessors to establish their shared ideal. The more difficult it is to come to an agreement about an essay's value the larger the variance among the essay's different assessments. The point is that both essays and assessors contribute to assessment variance. For any particular group one can calculate an average of the variation of all its essay assessments. This average variance, to which both assessors and essays contribute, is an evaluation deviation index.

The evaluation deviation index can be used to construct an interval around an essay's average mark. Essay assessments could be expected to fall within this interval with a fairly high level of probability. Thus, in stead of using only two best-bet marks all assessments falling within an essay's evaluation deviation interval are used to calculate the essay's final mark. In our system four assessments could contribute to the final mark, three peer assessments and one self assessment. Those assessments that do not fall within the evaluation deviation interval are excluded from further contributing to an essay's final mark. This procedure limits the average assessment deviation associated with an essay's assessments. For example, in our procedure we reduced the average assessment deviation from 10,93 to 7,42 in one group of 358 students and from 12,77 to 8,55 in a different group of 366 students. This reduces the interval within which an essay's final mark could be expected to fall, producing a more reliable rank order of final marks. A reliable rank order allows one to use an essay's final mark to establish an index of essay standard . The essay standard is an index expressing an essay's value relative to the other essays in the group. In addition the evaluation deviation interval allows one to calculate an index of essay agreement. This index takes the spread of assessments into account. The number of assessments falling within in an essay's evaluation deviation interval indicates the degree to which evaluators agree about the essay's mark.

Apart from selecting the assessments to be used to determine an essay's final mark the evaluation deviation interval also serves to track evaluator behaviour. Each evaluator assesses four essays, those of three peers, and his/her own essay. This allows one to count how many of the assessments performed by a particular evaluator are included in the evaluation deviation intervals of the various essays. One can also determine whether an evaluator tends to overestimate or underestimate essay marks, and express this assessment characteristic as assessment standard. Evaluators are considered poor assessors if most of their assessments are not within evaluation deviation intervals, and if they do not display a specific assessment standard, because it means they seldom agree with their peers about an essay's mark. This index of evaluator behaviour is called assessment consistency. By giving more weight to marks assigned by consistent (high agreement) evaluators and less weight to marks assigned by inconsistent (low agreement) evaluators it is possible to calculate an essay's final mark as a weighted mean. Thus not all contributions to an essay's constructed value are considered equally useful.

A last index made possible by the evaluation deviation interval combines an evaluator's own essay standard with his/her assessment standard. This index is called evaluation congruency. It is based on the point of view that evaluators who perform well on their own essays (ie who maintain high essay standards) would also maintain high assessment standards and thus tend to underestimate their peers (ie be too strict in their evaluations). By the same token one could expect correspondence between low essay standard and low assessment standard. Evaluators who do not adhere to this pattern are considered low on evaluation congruency.

Although the evaluation deviation strategy allows more sophisticated definitions of the various indices described above, similar indices can be derived in a best bet approach. The advantages of the best bet strategy are that it is quick and easy and that it guarantees significant differences between essays when they are ranked. The evaluation deviation strategy is a more cumbersome procedure but it takes into account the maximum amount of available information when constructing essay values. Both strategies have the same objective, namely to reduce the ratio of within essay variance over between essay variance, allowing for increased reliability when it comes to ranking essays. One also hopes that these procedures help to ensure that essays are assigned their proper ranks. It is not easy to extract information concerning the properness of ranks, though. One needs an external point of reference to act as a criterion of properness, but in this way properness becomes a question of validity, and validity is heavily burdened with traditional ideas of absolute truth. Nevertheless, we have to risk the detour through external criteria and standards to prove the possibility of self organisation in the closed system of peer assessment. In other words we have to compare students' behaviour when captured in teacher orbits with their behaviour when denting away on their own in the learning field.

Summary of main findings

Examinations provide an opportunity to test the validity of peer assessment procedures against external criteria. Our examination paper includes a question in which students are required to read and evaluate a research proposal using an evaluation schedule. Students' ratings of the research proposal are compared with benchmark ratings. The benchmark ratings are compiled by subject specialists on the lecturing team. Four subject specialists have to assess the proposal independently and then discuss their differences. The benchmark ratings are finalised only after student responses to the question have been taken into account. Mining the correlations between the examination marks and the various peer assessment indices turned up interesting results. Unfortunately the nature of the information calls for a somewhat involved discussion, which would not serve the purpose of this paper well. A summary of the main findings, which avoid speculative details, are perhaps more useful.

1. The first point to be made is that both the best bet and evaluation deviation strategies were successful ways to breach closed group processes and to produce a more reliable rank order of essays within the group. Both strategies had the desired effect of significantly lowering the ratio of intra-essay variance to inter-essay variance. Weighting the essay mean scores on the basis of evaluation quality caused a further numerical reduction of this ratio, but the reduction was not significant. For example, the average assessment deviation associated with essay means were reduced from 7,42 to 6,64 and from 8,55 to 7,68 for the groups mentioned above.
2. Self assessment should form part of a peer assessment procedure as it seems to significantly increase assessment validity. Essay marks calculated as the average of three peer assessments did not correlate significantly ($r_{xy} = 0,15$; N = 282) with criterion (examination) results. However, the correlation ($r_{xy} = 0,29$; N = 264)) between self assessments and examination results was statistically highly significant. When essay marks were calculated as the average of three peer assessments plus a self assessment mark the correlation coefficient improved from 0,15 to 0,28. We are hoping that this result may be an indication of the internalisation of standards. Students do their self assessments when they act as peer evaluators, which offers them the opportunity to reflect on their own work in relation to those of others.
3. Once self assessments have been included in calculating essay marks, further tampering does not seem to improve the validity of the peer evaluated marks as measured against an external criterion. When self assessments were excluded from contributing to peer assessments, essay marks failed to correlate significantly with examination results. The correlation coefficient was 0,15. These assessments required the intervention of a best bet strategy to increase the correlation from 0,15 to 0,24. When self assessments were incorporated with peer assessments in determining essay marks the correlation between essay marks and examination results turned out to be 0,28. Introducing the average deviation strategy made no difference. In fact the correlation between essay marks and examination results turned out to be lower (0,26) although this cannot be considered a significant difference. However, this does not mean that one should not employ an approach such as the average deviation strategy. Although the group may not be effected in general a few individual cases may reap the benefit of this approach. The rank orders of the untouched marks and the selected marks are highly correlated (0,97) but not exactly the same. However, weighting the selected marks did not change the essay rank order. Thus the weighting procedure did not add to the reliability of the essay marks. It is important to remember that the weights were determined on the basis of evaluator quality.
4. This raises a question about the validity of the distinction between good and poor evaluators. Evaluators were rated on the basis of evaluation consistency and essay standard. In other words good evaluators would be those individuals who maintain high levels of consistency in evaluating others and who themselves have written essays that receive high marks. Empirical evidence seems to support this approach. Marks assigned by more consistent evaluators correlated significantly with the examination result (0,24), whereas marks assigned by less consistent evaluators did not show a significant correlation (0,06). The same pattern held for essay standard. Marks assigned by evaluators whose own essays were judged to be of higher quality correlated significantly with the examination result (0,25), whereas marks assigned by those who submitted essays of lesser quality did not correlate well with the examination result (0,10). Thus it seems possible to distinguish between good and bad evaluators on the basis of evaluation consistency and essay standard. However, as discussed above, weighting essay marks on the basis of evaluation standard does not seem to improve rank order reliability. We are planning to research this aspect of peer evaluation in more detail later this year by introducing a benchmark procedure against which evaluator qualities can be measured.
5. It is interesting to note that evaluation standard (the tendency to under-evaluate or over-evaluate) did not present the expected correlation pattern. It was thought that individuals who submitted high quality essays would display a tendency to under-evaluate their peers' work because they would apply higher standards, and similarly that those who submitted lower quality essays would

maintain lower standards and therefore tend to over-evaluate their peers' essays. There was, however, no correlation between essay standard and evaluation standard (0,02). Furthermore, when marks assigned by under-evaluators were compared with the examination result the correlation proved to be not significant, whereas a comparison between marks assigned by over-evaluators and the examination result yielded a low but significant correlation (0,20). The same correlation pattern was observed with regard to the congruency index, which was not surprising because evaluation standard and essay standard served to constitute the congruency index (whether an individual applied similar standards in writing his/her own essay and in evaluating his/her peers). One may choose to speculate about the meaning and implications of this correlation pattern, but it is clear that evaluation standard is not a straightforward measure of evaluation quality. It is possible that the evaluation standard index is more an indication of an evaluator's ability to apply an evaluation schedule than it is a reflection of the academic standard maintained by an evaluator.

Conclusion

I have indicated how new needs require new approaches to teaching research methodology. I have outlined a learning field model to trace the roots of a peer assessment procedure, designed to form an integral part of our teaching approach, and I have presented empirical information to support the viability of this kind of assessment. One should keep in mind though that the results came from ad hoc procedures and not from a purposefully designed experiment. Although the correlation coefficients offered in support of the argument were statistically significant one should be mindful of their size and not consider them particularly powerful. Their meaningfulness lies not in their individual contributions, but rather in the pattern they constitute in support of the theoretical model. These results should be collaborated in follow-up studies before being considered conclusive.

However, despite this cautionary note we are confident that peer assessment presents itself as a viable technique in the world of praxis, firstly because the correlations clearly constitute a supportive pattern and secondly because we investigated peer assessment in a worst case scenario. The context presents a worst case scenario due to the comprehensiveness and the abstract nature of the assessment domain. Essays submitted for assessment dealt with a variety of different topics in a variety of ways. Students could choose any research topic and could use any one of a number of methodological procedures grounded in different philosophical approaches. The aspects to be assessed and the assessment criteria to be used had to be generic and not bound to content. Such criteria are abstract and difficult to apply. That students were able to do so in a meaningful manner is in fact surprising.

Or, perhaps we find this surprising only because we are used to students captured in our orbits, and not quite trusting those who present themselves as partners in the learning field. But if we want to expose ourselves to a learning field of peer assessment we have no choice but to accept our students as equal partners in curving out the generic reality of this kind of assessment procedure.

References

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, pp 155-159.

CSD. (1997). *CSD workshop on Research Methodology Training.* Council for Scientific and Industrial Research (CSIR), Pretoria.

Davidson, J.D. & Rees-Mogg, W. (1998). *The sovereign individual.* London: Pan books.

Department of Labour, (1997). *Green paper on skills development strategy of economic and employment growth in South Africa*. Pretoria: Author.

Grundy, S, (1987). *Curriculum: Product or Praxis*. London/New York: Falmer Press.

Rickey, R.C. (1995). Trends in instructional design: Emerging theory-based models. *Performance Improvement Quarterly*, 8(3), pp 96-110.

Schwahn, C.J. & Spady, W.G. (1998). *Total Leaders: Applying the best future-focussed change strategies to education*. Arlington: American Association of School Administrators